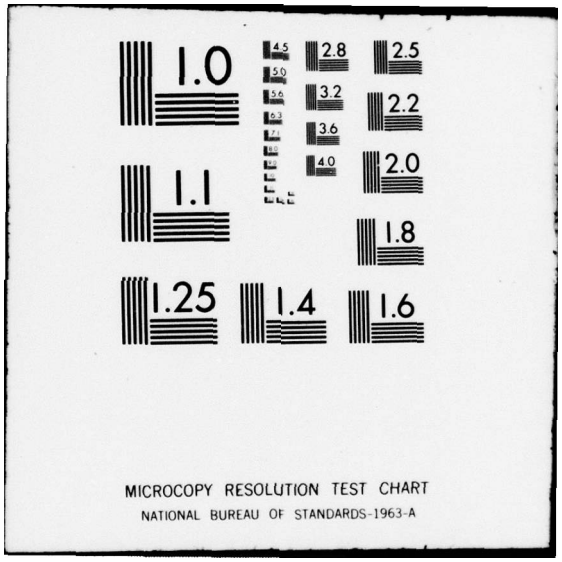
AD-A061 522 STANFORD UNIV CALIF DEPT OF OPERATIONS RESEARCH F/G 12/1
OPTIMALITY OF STATIONARY HALTING POLICIES AND FINITE TERMINATIO--ETC(U)
UNCLASSIFIED MAY 78 R E ERICKSON N00014-75-C-0493
TR-33 NL

END
DATE FILMED
1-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

LEVEL II

11

OPTIMALITY OF STATIONARY HALTING POLICIES AND FINITE TERMINATION OF SUCCESSIVE APPROXIMATIONS.

BY

RANEL E. ERICKSON

TR-33

TECHNICAL REPORT NO. 33

MAY 1978

Technical rept.

27p.

PREPARED UNDER
OFFICE OF NAVAL RESEARCH CONTRACT
N00014-75-C-0493 (NR-042-264)
NSF-ENG76-I2266

402 766

DEPARTMENT OF OPERATIONS RESEARCH
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

LEVEL II

11

OPTIMALITY OF STATIONARY HALTING POLICIES AND FINITE TERMINATION OF SUCCESSIVE APPROXIMATIONS

by

Ranel E. Erickson

Technical Report No. 33

May 1, 1978

Prepared Under
Office of Naval Research Contract N00014-75-C-0493*

Department of Operations Research
Stanford University
Stanford, California

*Also partially supported by National Science Foundation Grant ENG 76-12266.

OPTIMALITY OF STATIONARY HALTING POLICIES AND FINITE TERMINATION OF SUCCESSIVE APPROXIMATIONS

1. Introduction

Consider a discrete-time-parameter S-state finite-action branching Markov decision chain. Attention centers here on halting (resp., stopping) policies, i.e., those for which the expected population size at time N is zero for some N (resp., converges to zero as N approaches infinity). The value of a policy is the expected infinite-horizon income that it earns. The supremum of the values of the halting (resp., stopping) policies is the optimal halting (resp., stopping) value of the decision chain. In general these values are not the same. An optimal halting (resp., stopping) policy is one having maximum value in that class of policies.

Eaves and Veinott [1] have shown that if there is a stopping policy and all rewards are finite, then there is a stationary optimal stopping policy if and only if the optimal stopping value is finite. Moreover, when initiated with the value of a stopping policy, they have shown that the iterates of successive approximations converge to the optimal stopping value; also that value is a fixed point of the optimal return operator.

The purpose of this paper is to investigate the following additional problems under the hypothesis that the rewards are all real (resp., real or minus infinity) valued. When does there exist a halting stationary optimal stopping (resp., halting) policy? When do the iterates of successive approximations converge in finitely many steps assuming initiation with the value of a stationary halting policy?

The motivation for studying these problems comes from a companion paper with Veinott [2]. There we show that the problem of finding a minimum-concave-cost flow in a single-source network can be reduced to finding a stationary optimal halting policy in an associated branching Markov decision chain.

In order for there to be a halting optimal stopping (resp., halting) policy, there must be a halting policy. Section 3 is concerned with finding such a policy. To describe the results, let the halting time of a policy from a state s be the first time at which the expected population size is zero starting from s, if there is such a time; otherwise, let the halting time from s be infinity. Also let the halting time of a policy be the largest of its halting times from each state. The main result of Section 3 is that there is a stationary policy that simultaneously minimizes the halting time from each state, and each of the finite halting times of that policy is S or less. The proof of this result is a constructive combinatorial algorithm for finding the desired policy and its halting times from each state. One consequence of the above result is that there is a halting policy if and only if there is a stationary halting policy. Moreover, that is so if and only if the (stationary) policy found by the above algorithm is halting, or what is the same thing, its halting time is S or less. These results complement those of Rothblum [4, p. 74] concerning the case where instead every policy is halting (he calls halting policies nilpotent).

The main results are given in Section 4. There we characterize the existence of halting stationary optimal stopping (resp., halting)

policies by the condition that successive approximations terminates in finitely many steps. More precisely, suppose the rewards are real (resp., real or minus infinity) valued and successive approximations is initiated with the value of a stationary halting policy. Then the N-th iterate of successive approximations is a fixed point of the optimal return operator for some N if and only if that is so for some N not exceeding the largest of the halting times of the stationary halting policies; moreover, this occurs if and only if there exists a halting stationary optimal stopping (resp., halting) policy. Furthermore, when this is so, successive approximations terminates at the N-th iteration with such a policy, and its value is the indicated fixed point. Analogous results are also established where successive approximations is replaced by a Gauss-Seidel version. The running time of each of the above methods is proportional to the product of the numbers of states and nonzero data elements, i.e., rewards and transition probabilities.

2. Preliminaries.

Following Eaves and Veinott [1], Veinott [6], and Rothblum-Veinott [5] a branching Markov decision chain will now be described. Consider a population consisting of a finite set of individuals each of which is observed at a sequence of points in time labeled 1, 2, ... An individual observed at a given time point is found to be in a finite set $\mathcal{S}$ of <u>states</u> labeled 1, 2, ... , S. If there is no individual in any state, the population is said to have <u>stopped</u>. Each time an individual is observed in state s, an <u>action</u> a is chosen from a nonempty finite set A_s of possible actions in state s and a <u>reward</u> $-\infty < r(s,a) < \infty$ is received. The expected number of individuals in state t at time N+1 generated by each individual in state s at time N, given that action a was chosen at time N and given the states observed and actions taken at times 1, 2, ... , N-1 is assumed to be a real nonnegative function $p(t|s,a)$ depending only on t, s, and a.

Let $\Delta = X_{s=1}^{S} A_s$ be the set of all <u>decisions</u> and let a <u>policy</u> be a sequence $\pi = (\delta_1, \delta_2, \ldots)$ of decisions. Using a policy π means that if an individual is observed in state s at time N, then δ_N^s is the action chosen at that time. Let δ^∞ denote the sequence $(\delta, \delta, \ldots)$ and call it a <u>stationary policy</u>.

For any $\delta \in \Delta$, let r_δ be the S-vector whose s-th component is $r(s,\delta^s)$ and let P_δ be the nonnegative $S \times S$ matrix whose st-th element is $p(t|s,\delta^s)$. The elements of P_δ will be referred to as <u>transition rates</u>. Let $P_\pi^N \equiv P_{\delta_1} P_{\delta_2} \cdots P_{\delta_N}$ where $\pi = (\delta_1, \delta_2, \ldots)$. A state t is said to be <u>accessible</u> from state s in N steps using

policy π if $P^N_{\pi st} > 0$. A state is always accessible from itself in zero steps. A state is called <u>immediately accessible</u> from another when that is so in one step.

Define the S-vector v_π, called the <u>value</u> of π, by

$$v_\pi = \limsup_{k \to \infty} v^k_\pi \equiv \limsup_{k \to \infty} \sum_{N=0}^{k} P^N_\pi r_{\delta_{N+1}} .$$

Call the s-th component of v_π the <u>value</u> of π at initial state s. A policy π is called <u>transient</u> if $\sum_{N=0}^{\infty} P^N_\pi$ converges, and in this case

$$v_\pi = \sum_{N=0}^{\infty} P^N_\pi r_{\delta_{N+1}} ,$$

since the sum converges absolutely.

Call a policy π <u>halting</u> if $P^N_\pi = 0$ for some $N \geq 0$, and <u>stopping</u> if $P^N_\pi \to 0$ as $N \to \infty$. Of course halting policies are stopping. A halting (resp., stopping) policy π will be called <u>optimal halting</u> (resp., stopping) if $v_\pi \geq v_\sigma$ for all halting (resp., stopping) policies σ. In that event, v_π is called the <u>optimal halting</u> (resp., <u>stopping</u>) <u>value</u> of the system. The term "branching" refers to the fact that we require the transition rates to be nonnegative only. If in addition we assume that the sum of the transition rates in each row of P_δ is one or less for each decision δ, then we obtain an ordinary Markov decision chain. If such is the case, then the conditional probability that a subsystem enters the stopped state at time N+1 given that it is observed in state s at time N and action a is chosen then is $1 - \sum_{t=1}^{S} p(t|s,a)$. Associated with each decision δ is a (directed)

graph G_δ whose nodes are the states $1, 2, \ldots, S$ and whose arcs are the ordered pairs (s,t) such that $P_{\delta st} > 0$. A graph G is called circuitless if the nodes can be labeled $1, 2, \ldots, n$ so that if (s,t) is an arc in G, then $s < t$. A stationary policy δ^∞ is halting if and only if the graph G_δ is circuitless.

3. Characterization of Halting Policies.

Rothblum [4, p. 74] refers to halting policies as "nilpotent" policies. Define the halting time h_π of a policy π to be the smallest integer $N \geq 0$ such that $P_\pi^N = 0$ if π is halting, and set $h_\pi = \infty$ otherwise. Note that if policy π is used, then the individuals are almost surely in the stopped state at time h_π, i.e., the population has stopped.

A decision δ is called halting if that is so for δ^∞. Denote by Γ the set of halting decisions and let $h = \max_{\delta \in \Gamma} h_\delta$ be called the halting time of the system where $h_\delta = h_{\delta^\infty}$ and $h = \infty$ if $\Gamma = \emptyset$.

Define $h_{\pi t}$, the halting time of the policy π from state t, as the smallest integer N such that the t-th row of P_π^N vanishes if such an integer exists, and $h_{\pi t} = \infty$ otherwise. Thus $h_{\pi t}$ is the smallest integer $N \geq 0$ such that the population stops in N steps or less from t. Note that $h_{\pi t}$ is a "combinatorial" property of π in the sense that it depends on the location but not the magnitude of the positive elements of the P_δ.

The halting set $\mathcal{H}$ is the set of states t such that $h_{\pi t} < \infty$ for some π. Hence $t \in \mathcal{H}$ if there exists a policy with finite halting time from state t. The proof of the following result not only constructs

the set $\mathcal{H}$ but also exhibits a decision δ which satisfies $h_{\delta t} < \infty$ for each t in $\mathcal{H}$ where $h_{\delta t} \equiv h_{\delta^\infty t}$. Moreover, $h_{\delta t}$ is computed for each $t \in \mathcal{H}$.

Theorem 3.1. (Existence of Stationary Policies With Minimum Halting Times)

There is a stationary policy which simultaneously minimizes the halting times from each state. Moreover, the halting time of that policy is S or less from each state in $\mathcal{H}$.

Proof. Let H_k be the set of states from which there is a policy with halting time k or less. Then $H_o = \emptyset$. Also $H_k = H_{k-1} \cup I_k$ for $k \geq 1$ where I_k is the set of states s not in H_{k-1} such that for some action δ^s, say, in A_s, $p(t|s,\delta^s) = 0$ for each $t \notin H_{k-1}$. Since the I_k are disjoint, there is an integer $N \leq S$ such that $I_{N+1} = \emptyset$, and so $H_N = H_{N+1} = \cdots = \mathcal{H}$. For each $s \notin \mathcal{H}$ define δ^s in A_s arbitrarily. For each $s \in \mathcal{H}$, $h_{\delta s} = k$ where $s \in L_k$ and from the construction this is the minimum halting time from s. Q.E.D.

Note that the number of operations required to obtain a decision δ with minimum halting time is $O(S^3 z)$ where z is the average number of actions in a state. Moreover, if $\mathcal{H} = \mathcal{S}$, then the decision δ exhibited in the proof is halting. The proof implies that if $\mathcal{H} \neq \mathcal{S}$, then for each $i \in \mathcal{S} \setminus \mathcal{H}$ the i-th row of P_π^N contains a nonzero element for each policy π and each integer $N \geq 1$.

A matrix P is called nilpotent if $P^N = 0$ for some N. The spectrum of a matrix is defined as the set of its eigenvalues. Let

$L(G)$ be the number of nodes in a maximal chain (i.e., a directed path with no repetition of nodes) in the graph G. The following lemma states several known (e.g., Rothblum [4, p. 74], Kato [3, pp. 22, 38]) characterizations of halting decisions.

Lemma 3.2. (Characterization of Halting Decisions)

If δ is a decision, the following are equivalent:

(a) δ is halting.

(b) P_δ is nilpotent.

(c) The spectrum of P_δ is $\{0\}$.

(d) G_δ is circuitless.

(e) $h_\delta = L(G_\delta)$.

(f) $h_\delta \leq S$.

(g) $h_\delta < \infty$.

The next result is immediate from Theorem 3.1 and Lemma 3.2.

Theorem 3.3. (Existence of Halting Policies)

The following are equivalent:

(a) There exists a stationary halting policy.

(b) P_δ is nilpotent for some decision δ.

(c) The spectrum of P_δ is $\{0\}$ for some decision δ.

(d) G_δ is circuitless for some decision δ.

(e) $h_\delta = L(G_\delta)$ for some decision δ.

(f) $h_\delta \leq S$ for some decision δ.

(g) $h \leq S$.

(h) There exists a halting policy.

(i) $\mathcal{H} = \mathcal{S}$.

Remark. The computation of h is NP-complete since a solution can be verified in polynomial time and the longest path problem (which is NP-complete) can be transformed to the present problem in polynomial time. To see this, let the states be nodes and let the actions determine whether to stop or choose an arc leading to an adjacent node. The graph of a halting decision δ is circuitless and h_δ is the length of it's longest path. Note that h_δ equals the number of nodes if and only if there is a Hamiltonian path (i.e., a path containing all the nodes) in the graph.

4. Stopping Optimality and Finite Termination of Successive Approximations.

In this section the existence of a halting stationary optimal halting (resp., stopping) policy will be characterized by the condition that successive approximations terminates in finitely many steps. First, define the optimal return operator $\mathcal{R}$ by $\mathcal{R}v = \max_{\delta \in \Delta} R_\delta v$, where $R_\delta v = r_\delta + P_\delta v$. The method of successive approximations is the repeated application of the optimal return operator, $\mathcal{R}^k v$ being the k-th approximation using v as the initial approximation. When $r_\delta \gg -\infty$ for all δ, Eaves and Veinott [1] have shown that for every stopping value v^o, $\mathcal{R}^k v^o \uparrow v^*$ where v^* is the optimal stopping value and there is a stationary optimal stopping policy if and only if v^* is finite. In our study of halting policies, the initial approximation v^o will be the value of any halting decision γ where we require only that $-\infty \leq r_\gamma \ll \infty$. Then v_γ is the unique $v \ll \infty$ satisfying the recursive system $v = R_\gamma v$. Recall also that a halting γ can be constructed as in the proof of Theorem 3.1, if one exists.

Lemma 4.1.

If γ is a halting decision, then $\mathcal{R}^k v_\gamma$ is nondecreasing in $k \geq 0$ and is the value of a halting policy for each $k \geq 0$. Also, $\mathcal{R}^k v \geq R_\gamma^k v = v_\gamma$ for each $k \geq h_\gamma$ and v.

Proof.

Since γ is halting $\mathcal{R} v_\gamma \geq R_\gamma v_\gamma = v_\gamma$ so $\mathcal{R}^{k+1} v_\gamma \geq \mathcal{R}^k v_\gamma$ for $k \geq 0$. Also $\mathcal{R}^k v_\gamma$ is the value of a halting policy that uses

γ in each period following period k. The final assertion is immediate on noting that $P_\gamma^{h_\gamma} = 0$.

Under the assumption that $r_\delta >> -\infty$ for all δ, Eaves and Veinott [1] have characterized when a stationary optimal stopping policy exists. By contrast, the next result characterizes when that policy can also be taken to be halting. It asserts that such a policy exists if and only if the method of successive approximations terminates in finitely many steps when initiated with the value of any halting decision.

Theorem 4.2. (Existence of Stationary Optimal Halting Policies)

If there is a halting policy, then the following are equivalent.

(a) There is a stationary optimal halting policy.

(b) $\mathcal{A}^i v_\gamma$ is the optimal halting value for every $i \geq h$ (resp., some $i \geq 0$) and every (resp., some) halting decision γ.

(c) $\mathcal{A}^i v_\gamma$ is a fixed point of $\mathcal{A}$ for every $i \geq h$ (resp., some $i \geq 0$) and every (resp., some) halting decision γ.

If also $r_\delta >> -\infty$ for all δ, then the above conditions remain equivalent on replacing "optimal halting" with "optimal stopping" and inserting "halting" before "stationary" everywhere.

Proof.

(a) $\Longrightarrow$ (b). By hypothesis, there is a stationary optimal halting policy δ^∞, say. For each halting decision γ and $i \geq h$ it follows from Lemma 4.1 that $v_\delta \geq \mathcal{A}^i v_\gamma \geq R_\delta^i v_\gamma = v_\delta$ so $v_\delta = \mathcal{A}^i v_\gamma$.

(b) => (c). By Lemma 4.1 and hypothesis $\mathcal{R}^{i+1}v_\gamma \geq \mathcal{R}^i v_\gamma \geq \mathcal{R}^{i+1}v_\gamma$ so $\mathcal{R}(\mathcal{R}^i v_\gamma) = \mathcal{R}^i v_\gamma$.

(c) => (a). By Lemma 4.1 $v^k \equiv \mathcal{R}^k v_\gamma$ is nondecreasing in $k \geq 0$, and by hypothesis $v^k = v^i$ for $k \geq i$. Choose the decisions $\delta_0 = \gamma$, δ_1, δ_2, ... recursively so that $v^k = R_{\delta_k} v^{k-1}$ and $\delta_k^s = \delta_{k-1}^s$ if $v_s^k = (R_{\delta_{k-1}} v^{k-1})_s$ for $s \in S$ and $k > 0$. We now show that δ_i is halting. If not, the graph of δ_i contains a circuit Γ on the set of states C say. Let $m \geq 0$ be the smallest integer such that $\delta_m^s = \delta_{m+1}^s = \cdots = \delta_i^s$ for each $s \in C$. Since $\delta_0 = \gamma$ is halting, $m > 0$. Because $v^k \geq v^{k-1}$ (with $v^{-1} \equiv v_\gamma$)

$$(1) \qquad v_s^{k+1} \geq (R_{\delta_k} v^k)_s \geq (R_{\delta_k} v^{k-1})_s = v_s^k$$

for each $s \in \mathcal{S}$ and $k \geq 0$.

Now we show by induction on k that for each $m-1 \leq k \leq i$, at least one of the inequalities in (1) is strict for some $s \in C$ (depending on k). Since $\delta_{m-1}^s \neq \delta_m^s$ for some $s \in C$, the first inequality in (1) is strict for that s by construction so the claim holds for $k = m-1$. Suppose it holds for $k-1 (m-1 \leq k-1 < i)$ and consider k. Then $v_t^k > v_t^{k-1}$ for some $t \in C$. Now by definition of C, there is an $s \in C$ such that (s,t) is an arc of the graph of δ_k. Thus the second inequality in (1) is strict if $v_u^k > -\infty$ for each u in the set $\mathcal{U}$ of states immediately accessible from s when using δ_k. Since $v^k \geq v^m$, it suffices to show $v_u^m > -\infty$ for each $u \in \mathcal{U}$. We now show that this is the case.

Recall that $v_t^m > -\infty$ for some $t \in C$. Put $\delta = \delta_m$. Since $\delta_k^s = \delta^s$ and the graph G_δ contains a chain (in Γ) from t to s

having N, say, nodes, then $(P_\delta^N)_{tu} > 0$ for each $u \in \mathcal{U}$. Also because $R_\delta v^m \geq v^m$ by (1), $R_\delta^N 0 + P_\delta^N v^m = R_\delta^N v^m \geq v^m$. Hence $v_u^m > -\infty$ for each $u \in \mathcal{U}$ as claimed, so the second inequality in (1) is strict.

From the above, $v^{i+1} > v^i$, which is a contradiction. Thus, δ_i is halting. Also by (1) and $v^{i+1} = v^i$, $v^i = R_{\delta_i} v^i$ and so $v^i = v_{\delta_i}$.

Hence as in Eaves and Veinott [1], $v^i \geq R_\delta v^i$ for all δ, so on iterating this inequality using the decisions in a policy π we get $v^i \geq v_\pi^N + P_\pi^N v^i$ for $N \geq 1$. Thus if π is halting, $v^i \geq v_\pi$. Hence δ_i^∞ is a stationary optimal halting policy, establishing (a).

To complete the proof, it remains to note that if $r_\delta \gg -\infty$ for all δ, "optimal halting" is replaced by "optimal stopping" and "halting" is inserted before "stationary" everywhere in the above theorem and its proof, and "π is halting" is replaced by "π is stopping" in the preceding paragraph, then the proof is valid as is. (The finiteness of the r_δ's is used only in the next to last sentence of the proof that (c) implies (a).)

Remark. One immediate consequence of the above result is that if δ^∞ is a halting stationary optimal halting (resp., stopping) policy, then $\mathcal{R}^{h_\delta} v_\gamma$ is the optimal halting (resp., stopping) value for each halting decision γ.

Algorithm. The above proof justifies the following procedure for determining whether or not a halting stationary optimal halting (resp., stopping) policy exists, and if so, producing such a policy. First, use the construction of Theorem 3.1 to determine whether or not a halting decision exists, and if so, find one, say γ. In the latter event construct $\delta_0, \delta_1, \delta_2, \ldots$ inductively as in the proof of Theorem 4.2 until an $0 \leq i \leq S$ is found for which $\mathcal{R}^i v_\gamma$ is a fixed point

of $\mathcal{A}$. In that event δ_i^∞ is the desired halting stationary optimal halting (resp., stopping) policy. If no halting decision γ exists, or if γ is halting but $\mathcal{A}^i v_\gamma$ is not a fixed point for any $0 \leq i \leq S$, then no halting stationary optimal (resp., stopping) policy exists.

Theorem 4.2 implies that if the r_δ's are all finite, then the optimal halting and stopping values coincide. However, this need not be the case when $r_{\delta s} = -\infty$ for some δ and s as the following example illustrates.

Example. The Optimal Halting and Stopping Values Need Not Coincide.

Suppose there is one state and two actions γ and δ are available therein. If γ is chosen at time k, the population receives a reward of $-\infty$ and stops. If δ is chosen at time k, the population receives zero reward, remains in the state with probability 1/2, and stops with probability 1/2. Both $v_\gamma = -\infty$ and $v_\delta = 0$ are fixed points of $\mathcal{A}$ where $\mathcal{A}v = \max(1/2\ v, -\infty)$. Also v_δ is the optimal stopping value and v_γ is the optimal halting value.

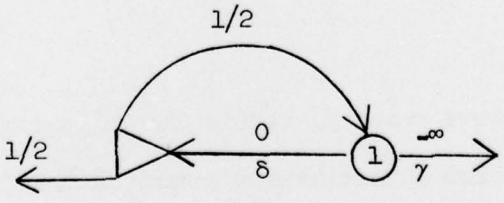


Figure 1.

Recall z is the average number of actions in a given state. Note that S^2z is the number of additions and approximately the number of

comparisons required in one application of the optimal return operator. Thus the number of operations needed in S steps of successive approximations is $O(S^3z)$.

Incidentally not all decisions $\delta_0, \delta_1, \delta_2, \ldots$ chosen above are necessarily halting even though the first and the last decisions are halting.

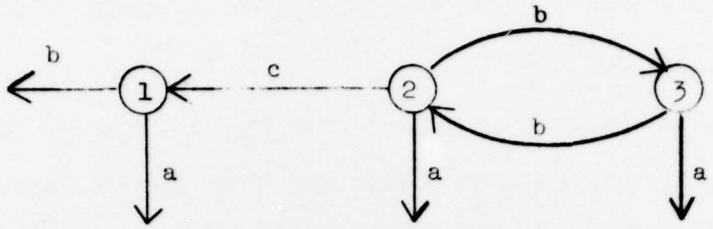
Example. Halting and Nonhalting Decisions are Obtained During Intermediate Steps.

Referring to Figure 2 below, consider a system with three states labeled 1, 2, 3. There are two actions available in states 1 and 3 while state 2 has three actions. Let a, b, c denote respectively the first, second, and third actions available in a state. Let the rewards be $r(1,a) = 0$, $r(1,b) = r(2,a) = r(2,b) = r(2,c) = r(3,a) = r(3,b) = 1$. Let the transition rates be $p(1|2,c) = 1$, $p(2|3,b) = p(3|2,b) = 1/2$ and all others zero. Note that $\delta_0 = (a,a,a)$ is halting, but $\delta_2 = (b,b,b)$ is not.

The Gauss-Seidel Method of Successive Approximations.

In general the method of successive approximations can be improved by updating the value v_s^{k+1} immediately after it is computed. Define $(T_{s\delta}v)_k = (R_\delta v)_s$ if $k = s$, and $(T_{s\delta}v)_k = v_k$ otherwise. Let $T_\delta = T_{S\delta} \cdots T_{1\delta}$. Observe that $T_\delta v = r_\delta + Q_\delta v$ where $Q_\delta = P_{s\delta} \cdots P_{1\delta}$ and $P_{s\delta}$ is formed by replacing the s^{th} row of the identity matrix by the s^{th} row of P_δ. Define the Gauss-Seidel optimal return operator $\mathfrak{T}$ by $\mathfrak{T}v = \max_{\delta \in \Delta} T_\delta v$.

Our goal now is to establish analogs of Lemma 4.1 and Theorem 4.2 for the Gauss-Seidel optimal return operator. An elegant way to show this is to note that both of the above results and their proofs are



n	0	1	2	3	4
v^n	0	1	2	2	2
	1	1.5	2	2	2
	1	1.5	1.75	2	2
δ_n	(a,a,a)	(b,b,b)	(b,b,b)	(b,c,b)	(b,c,b)

Figure 2.

valid as is when $\mathcal{R}$ and R_δ are replaced respectively by $\mathcal{T}$ and T_δ. However, there is one complication in this approach. It is that $\mathcal{T}^k v_\gamma$ is not the value of a halting policy in the sense defined here. However, $\mathcal{T}^k v_\gamma$ is the value of a generalized halting policy in which the action chosen at each time depends not only on the state occupied at that time but also on the last state visited. More precisely, let $\delta_1, \ldots, \delta_k$ be decisions with $\mathcal{T}^k v_\gamma = T_{\delta_1} \cdots T_{\delta_k} v_\gamma$. Then $\mathcal{T}^k v_\gamma$ is the value of the generalized halting policy in which one uses $\delta_1, \ldots, \delta_k$ consecutively, each for as long as the states visited decrease strictly. Thus if δ_i is used at time n and the system is in states s and t at times n and $n+1$ respectively, then one uses δ_i in period $n+1$ if $s > t$ and δ_{i+1} in that period otherwise.

We can avoid the use of these generalized halting policies by proving a variant of the natural analog of Lemma 4.1. First, an upturn

in a sequence of N positive integers $s_1, s_2, \ldots, s_N$ is an integer $1 \leq n \leq N$ such that $s_{n-1} \leq s_n$ where $s_0 = 0$. Let $g_{\pi s}$ be the limit as $N \to \infty$ of the maximum number of upturns in the first N states visited starting with $s_1 = s$ and using π among those sequences of length N having positive probability. Let $g_\pi \equiv \max_{s \in S} g_{\pi s}$ be called the upturn number of the policy π. Also let $g_\delta \equiv g_{\delta^\infty}$ and $g_{\delta s} = g_{\delta^\infty s}$.

Given a stationary policy δ^∞, we can compute g_δ by considering the graph G_δ. Clearly G_δ is circuitless if and only if $g_\delta < \infty$, since the existence of a circuit implies there is at least one arc (i,j) in G_δ, called an upturn arc, such that $i \leq j$. If G_δ is circuitless, then g_δ is one plus the number of upturn arcs in the chain that has the maximum number of upturn arcs. This chain is easily computed by any algorithm for finding a maximum cost chain when a value of one is assigned to each upturn arc and all others are assigned value zero. An analog of Lemma 4.1 is now presented.

Lemma 4.3.

If γ is a halting decision, then $\mathcal{F}^k v_\gamma$ is nondecreasing in $k \geq 0$. Also $\mathcal{R}^k v \leq \mathcal{F}^k v \leq \mathcal{R}^{kS} v$ for all $k \geq 0$ and all v such that $v \leq \mathcal{R} v$. Moreover, for each v, $\mathcal{F}^k v \geq T_\gamma^k v$ for all $k \geq 0$ and $T_\gamma^k v = v_\gamma$ for all $k \geq g_\gamma$.

Proof.

Since γ is halting, $\mathcal{F} v_\gamma \geq T_\gamma v_\gamma = v_\gamma$. Thus $\mathcal{F}^{k+1} v_\gamma \geq \mathcal{F}^k v_\gamma$ for all $k \geq 0$.

Let w be such that $w \leq \mathcal{R} w$ and let δ be a decision such that $\mathcal{R} w = R_\delta w$. Thus $T_{t\delta} w \geq w$ for all t. Iterating this inequality,

$w^s \equiv T_{s-1,\delta} \cdots T_{1\delta} w \geq w$ for $s = 2, 3, \ldots, S$. Set $w^1 \equiv w$. For $s = 1, 2, \ldots, S$ we have $(\mathcal{R}w)_s = (R_\delta w)_s \leq (R_\delta w^s)_s = (T_\delta w)_s \leq (\mathcal{T}w)_s$, where the first inequality follows from the monotonicity of R_δ and the second equality follows from the definition of T_δ. Thus $\mathcal{R}w \leq \mathcal{T}w$ for all w such that $w \leq \mathcal{R}w$. Hence since $v \leq \mathcal{R}v$, we have by induction that

$$\mathcal{R}^k v \leq \mathcal{T}\mathcal{R}^{k-1} v \leq \mathcal{T}^k v \tag{1}$$

where the first inequality follows from what was shown above on letting $w = \mathcal{R}^{k-1} v$ and the second inequality follows from the monotonicity of $\mathcal{T}$ together with the induction hypothesis.

From the definition of $T_{s\alpha}$,

$$w \leq \mathcal{R}w \text{ implies } T_{s\alpha} w \leq \mathcal{R}w \text{ for all } s \text{ and } \alpha . \tag{2}$$

Let y be such that $y \leq \mathcal{R}y$ and let β be a decision satisfying $\mathcal{T}y = T_\beta y$. Then (2) implies $T_{1\beta} y \leq \mathcal{R}y$. Applying $T_{s\beta}$ sequentially for $s = 2, 3, \ldots, S$ yields

$$\mathcal{T}y = T_\beta y \leq T_{s\beta} \cdots T_{s+1,\beta} \mathcal{R}^s y \leq \mathcal{R}^S y \tag{3}$$

where the first inequality follows by iteratively applying (2) for $j = 1, 2, \ldots, s$ on setting $w = \mathcal{R}^j y$. Moreover by induction

$$\mathcal{T}^k v \leq \mathcal{T}\mathcal{R}^{S(k-1)} v \leq \mathcal{R}^{Sk} v \tag{4}$$

where the first inequality is true from the induction hypothesis together with the monotonicity of $\mathcal{T}$ and the second inequality follows from (3) on letting $y = \mathcal{R}^{S(k-1)} v$. We conclude from (1) and (4) that $v \leq \mathcal{R}v$

implies $\mathcal{R}^k v \leq \mathcal{F}^k v \leq \mathcal{R}^{Sk} v$ for all $k \geq 0$.

From the definition of $\mathcal{F}$, $\mathcal{F} w \geq T_\gamma w$ for all w. By induction, $\mathcal{F}^k v \geq \mathcal{F} T_\gamma^{k-1} v \geq T_\gamma^k v$ where the first inequality follows from the monotonicity of $\mathcal{F}$ together with the induction hypothesis and the second inequality follows from what was shown above on setting $w = T_\gamma^{k-1} v$. Thus $\mathcal{F}^k v \geq T_\gamma^k v$ for all k.

It remains to show $T_\gamma^k v = v_\gamma$ for $k \geq g_\gamma$. Call a state that is immediately accessible from a state s when γ is used a <u>follower</u> of s. Let $J_k \equiv \{s: g_{\gamma s} \leq k\}$ and $L_w \equiv \{s: w_s = v_{\gamma s}\}$ where w is an S-vector. If the followers of s are in L_w, then

$$(5) \qquad (T_{s\gamma} w)_s = r_{\gamma s} + (P_\gamma w)_s = r_{\gamma s} + (P_\gamma v_\gamma)_s = v_{\gamma s} .$$

It is sufficient to show

$$(6) \qquad J_k \subseteq L_{T_\gamma^k v} \quad \text{for} \quad k = 0, 1, \ldots$$

since $J_k = \{1, 2, \ldots, S\}$ for $k \geq g_\gamma$. For $k = 0$, (6) becomes $\emptyset \subseteq L_v$ which is trivially true. By induction, assume (6) holds for k-1 and consider k. Let $w^j \equiv T_{j\gamma} \cdots T_{1\gamma} T_\gamma^{k-1} v$ for $1 \leq j \leq S$ and $w^o \equiv T_\gamma^{k-1} v$. For simplicity of notation, let $L_j \equiv L_{w^j}$. It is sufficient to show that

$$(7) \qquad J_k \cap \{1, 2, \ldots, j\} \subseteq L_j \quad \text{for} \quad j = 0, \ldots, S$$

since $w^S = T_\gamma^k v$ and, when $j = S$, (7) becomes (6). Now (7) is trivial for $j = 0$. Suppose it holds for j-1 and consider j. Since $T_{j\gamma}$ changes only the j-th component of any vector,

$$(8) \qquad L_{j-1} \subseteq L_j \cup \{j\} .$$

Thus if j is not in J_k, then by the induction hypothesis $J_k \cap \{1, \ldots, j\} \subseteq J_k \cap \{1, \ldots, j-1\} \subseteq L_j$. If j is in J_k, then the set B_j of followers of j that are in $H_j \equiv \{j+1, j+2, \ldots, S\}$ must have upturn number at most $k-1$ and the set E_j of followers of j that are in $\{1, 2, \ldots, j-1\}$ must have upturn number at most k. From (6) for $k-1$, $B_j \subseteq J_{k-1} \cap H_j \subseteq L_o \cap H_j \subseteq L_{j-1}$ where the third inclusion follows from the fact that the operators $T_{p\gamma}$ for $1 \leq p \leq j$ do not change the components of any vector with indices in H_j. From (7) for $j-1$, $E_j \subseteq L_{j-1}$. Thus the followers of j are in L_{j-1}, and from (5) $w_j^j = (T_{j\gamma} w^{j-1})_j = v_{\gamma j}$. Hence $j \in L_j$ and thus by (7) for $j-1$ and (8), (7) holds for j which completes the proof.

Remark. Lemma 4.3 implies that the Gauss-Seidel method converges faster than the usual method of successive approximations, i.e., the value of each Gauss-Seidel iterate is always greater than that of the usual method.

Now using the above lemma we obtain the analog of Theorem 4.2.

Theorem 4.4.

Theorem 4.2 remains valid if $\mathcal{R}$ is replaced by $\mathcal{T}$ and h by g.

Proof.

The proof is identical to that of Theorem 4.2 where $\mathcal{R}$, R_γ, P_γ, and h are replaced by $\mathcal{T}$, T_γ, Q_γ, and g respectively and Lemma 4.3 is used instead of Lemma 4.1 with only one exception. The exception is that the above replacements are not made in the last paragraph of the proof that (c) implies (a). Q.E.D.

Remark 1. The remark and algorithm following Theorem 4.2 also remain valid with the obvious replacements.

Remark 2. The preceding theorem implies that if the states are labeled such that P_δ is lower triangular with zero diagonal elements for some optimal halting (resp., stopping) policy δ^∞, then $\mathfrak{R} v_\gamma = v_\delta$ for every halting decision γ since $g_\delta = 1$. The number of operations is thus reduced by a factor of g.

In some cases it is possible to determine whether there exists a halting optimal stopping policy without computation. The following theorem gives sufficient conditions for the existence of a halting stationary optimal stopping policy.

Theorem 4.5.

Assume there exists a halting policy and $-\infty \ll r_\delta \leq 0$ for each decision δ. If for each decision δ and each circuit C in G_δ, the product of the transition rates around C is one or more and $r_{\delta t} < 0$ for some $t \in C$, then there exists a halting stationary optimal stopping policy.

Proof.

By hypothesis, there exists a halting policy and hence a stopping policy. Also $0 \geq \mathfrak{R} 0$, so by a result of Eaves and Veinott [1], there exists a stationary optimal stopping policy δ^∞, $v_\delta \leq 0$, and $\mathfrak{R} v_\delta = v_\delta$. Suppose G_δ has a circuit consisting of the nodes $\{1, 2, \ldots, m\}$ with $r_{\delta 1} < 0$, say. Then $v_{\delta 1} < p_1 \cdots p_m v_{\delta 1} \leq v_{\delta 1}$ where $p_k = p(k+1|k,\delta^k)$ for $k < m$ and $p_m = p(1|m,\delta^m)$. This is a contradiction, so δ is halting and the theorem follows.

Acknowledgments.

This work was motivated by the study of minimum-concave-cost network flows discussed in a companion paper [1] co-authored with Professor Arthur F. Veinott, Jr. His influence was deeply felt and appreciated in the present paper also. I sincerely thank him for suggesting this problem, guiding the development of the theory, simplifying the proofs and notation, suggesting terminology, carefully correcting and improving the exposition, and suggesting a number of results and proofs. I am also grateful to him and to Professor B. Curtis Eaves for making available their unpublished work on optimal stopping policies which is the foundation on which this paper builds.

REFERENCES

[1] Eaves, B. C. and A. F. Veinott, Jr. (forthcoming). Maximizing Values and Stopping Values of Markov Decision Chains by Policy Improvement.

[2] Erickson, R. E. and A. F. Veinott, Jr. (forthcoming). Minimum-Concave-Cost Single-Source Network Flows.

[3] Kato, T. (1976). Perturbation Theory for Linear Operators, 2nd ed. Springer-Verlag, New York, N. Y.

[4] Rothblum, U. (1974). Multiplicative Markov Decision Chains. Doctoral Dissertation, Operations Research Dept., Stanford University, Stanford, California.

[5] Rothblum, U. and A. F. Veinott, Jr. (forthcoming). Average-Cumulative-Optimality in Polynomially-Bounded Branching Markov Decision Chains.

[6] Veinott, A. F., Jr. (1974). Markov Decision Chains. Studies in Optimization 10, in MAA Studies in Mathematics, 124-159.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 33	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) OPTIMALITY OF STATIONARY HALTING POLICIES AND FINITE TERMINATION OF SUCCESSIVE APPROXIMATIONS		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) RANEL E. ERICKSON		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0493
9. PERFORMING ORGANIZATION NAME AND ADDRESS DEPARTMENT OF OPERATIONS RESEARCH STANFORD UNIVERSITY _ STANFORD, CA.		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (NR-042-264)
11. CONTROLLING OFFICE NAME AND ADDRESS LOGISTICS & MATHEMATICAL STATISTICS BRANCH OFFICE OF NAVAL RESEARCH ARLINGTON, VIRGINIA		12. REPORT DATE May 1, 1978
		13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale.
Its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Also partially supported by National Science Foundation Grant ENG-76-12266,

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Markov decision chains, Stopping policies, Halting policies,
Successive approximations, Gauss-Seidel method

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See Reverse side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT

The stopping and halting optimality of stationary halting policies in discrete-time-parameter S-state finite-action branching Markov decision chains are characterized by the finite termination of successive approximations. A policy is called halting (resp., stopping) if the expected population size at time N is zero for some N (resp., converges to zero as N approaches infinity). The value of a policy is the expected infinite-horizon income that it earns. An optimal stopping (resp., halting) policy is one having maximum value in that class of policies. It is shown that when the rewards are real (resp., real or minus infinity) valued, the N-th iterate of successive approximations (and a Gauss-Seidel improvement thereof) is a fixed point of the optimal return operator for some N when initiated with the value of a stationary halting policy if and only if that is so for some $N \leq S$; moreover this occurs if and only if there exists a halting stationary optimal stopping (resp., halting) policy. Furthermore, when this is so, successive approximations (and its Gauss-Seidel improvement) terminates at the N-th iteration with such a policy, and its value is the indicated fixed point. A combinatorial algorithm for finding a stationary halting policy or showing one does not exist is given. The running time of each of the above algorithms is proportional to the product of the numbers of states and nonzero transition probabilities. The results are applied in a companion paper with Veinott to find minimum-concave-cost flows in single-source networks.